The CDF Computing and Analysis System: First Experience

Rick COLOMBO[1], Paul HUBBARD[1], Stephan LAMMEL[1], Mark LEININGER[1], Fedor RATNIKOV[2], Terry WATTS[2] for CDF Offline Group
[1](Fermi National Accelerator Laboratory, Batavia, Illinois 60510, USA)
[2](Rutgers University Physics Dept, Piscataway, New Jersey 08854, USA)

Abstract

The Collider Detector at Fermilab (CDF) collaboration records and analyses proton anti-proton interactions with a center-of-mass energy of 2 TeV at the Tevatron. A new collider run, Run II, of the Tevatron started in April. During its more than two year duration the CDF experiment expects to record about 1 PetaByte of data.

With its multi-purpose detector and center-of-mass energy at the frontier, the experimental program is large and versatile. The over 500 scientists of CDF will engage in searches for new particles, like the Higgs boson or supersymmetric particles, precision measurement of electroweak parameters, like the mass of the W boson, measurement of top quark parameters, and a large spectrum of B physics.

The experiment has taken data and analysed them in previous runs. For Run II, however, the computing model was changed to incorporate new methodologies, the file format switched, and both data handling and analysis system redesigned to cope with the increased demands.

This paper (4-036 at Chep 2001) gives an overview of the CDF Run II compute system with emphasize on areas where the current system does not match initial estimates and projections. For the data handling and analysis system a more detailed description is given.

Keywords: compute system, analysis model, experimental software, data analysis, data handling, analysis system, CDF, Run II

1 Introduction

The Collider Detector at Fermilab (CDF) [1] is a large multi-purpose particle detector at the Fermilab Tevatron. With the upgrade of the accelerator complex at Fermilab for Run II, anti-proton interactions are delivered at a center-of-mass energy of $\sqrt{s} = 2\,\mathrm{TeV}$ with expected luminosity up to $2 \cdot 10^{32}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. Beam crossing time is reduced to 396 ns. Collisions started in April 2001 and the detector has been recording, and the collaboration analysing, data since that date.

To cope with the accelerator upgrades, the detector had a major upgrade during the same period. And a similar upgrade to the compute and analysis systems[2] was also needed. The procedural FORTRAN 77 environment of Run I was replaced with an object oriented C++ environment. The format of the data files was changed from YBOS to ROOT (to enable more direct access in the future). The software development and build tools are now based on the SoftRelTools package to leverage expertise from other HEP experiments. The data handling system was redesigned to overcome some of the shortcomings encountered in Run I and to cope with the increased data and analysis demands. Since the central analysis system provides most of the experiment's analysis power, it has to be upgradable beyond the current run. Redesigned systems incorporate new computing technologies and build on the successful Run I strategy.

2 Software Development

The code management in CDF is based on the Concurrent Version System (CVS) package. The package provides software version control for the source code of the experiment. It is used in

a client-server mode to facilitate software development in the highly distributed environment of the collaboration. For release management and library/binary building the SoftRelTools package, which was originally designed by BaBar, is being used. Fermilab's UNIX Product Support (UPS) and UNIX Product Distribution (UPD) products are being used to access and distribute the software packages.

There are currently over 250 packages in the CDF reconstruction and analysis environment. While the development version of all packages is updated nightly, a so called "frozen" release including most of the packages is made about once a month. Such releases are validated in great detail and assembling them is a manpower intensive operation. To satisfy the need of the collaboration for a more up-to-date release (than the frozen releases) and more reliable (than the development release), a weekly integration release was introduced earlier this year. This release requires a successful build of all executables but no validation.

CDF uses the KAI C++ compiler from Kuck and Associates. KAI translates C++ into ANSI C that can then be compiled using the native compilers. The product was initially selected because it was closest to the ISO C++ standard and available on all three compute platforms of the experiment. CDF code is required to be standards compliant. A change to compiler from KAI to gcc is being evaluated.

3 Analysis Framework

The CDF experiment decided for an object oriented (OO) infrastructure with a modular program architecture for the analysis framework. The new analysis driver (AC++) was co-developed with the BaBar experiment. The idea is similar to the analysis driver used in the Run I FORTRAN environment. The framework allows the user to configure amd control the execution of selected modules at run time. Reconstruction (or analysis) units form modules which are controlled through the analysis driver. Modules are autonomous and do not communicate directly with each other.

In addition to module configuration and execution control, the analysis framework provides the infrastructure to manage the data flow, provides error handling, recovery, and abort mechanisms, and accumulates execution and event selection statistics.

Dedicated input and output modules interface to the data handling system to provide a convenient, user friendly interface[3]. With those modules, users can access data by dataset name and need no longer keep track of file lists and file locations.

The analysis driver with all reconstruction modules linked in the current default debug mode to an executable over 230 MByte in file size on the main machine in the central analysis cluster running IRIX 6.5. An analysis executable containing only the modules to dump the data objects and banks reaches 150 MByte.

4 Data Persistency

In Run I CDF stored all its data in YBOS files. Data banks of the events were encapsulated into logical records and stored sequentially inside such files. For Run II the ROOT format was selected to store the CDF data[4]. With the switch to this format, a new fully OO event data model (EDM) was developed. Right now data are contained in one ROOT branch. The EDM was implemented during the last three years and is now used by all reconstruction and analysis modules.

The goal of the EDM design was to separate the storage solution, ROOT, from the data objects seen inside the analysis and reconstruction modules. Two classes of objects can be used to create data that can be saved in the event record: StorableObjects, which can be stored as a stand alone objects and StreamableObjects which may only be constituents of another

storable or streamable object. In the design of objects, both optimization for use in analysis and reconstruction modules, and for high performance input and output is important.

The speed at which data of the experiment can be read and restored into objects with the debug executables that are currently being used is at a few MByte/s on the main central analysis machine, the SGI Origin 2000. Work on the performance is ongoing and the people working on it are confident to reach 12 MByte/s. Work to split an event into multiple ROOT branches is in progress.

5 Data Handling

Both raw and reconstructed data are stored on tape. User analysis jobs and reconstruction farm jobs access data only from disk. A staging and disk inventory management system (DIM/Stager)[5] moves data between tape and disk in parallel with the data processing from disk.

The tape archive is an automated tape library from ADIC model AML/2. The tape drives in the library are directly attached to compute and data serving nodes via parallel SCSI. Each compute or data node thus runs a DIM/Stager system and all access to the tape library is by such systems. One such DIM/Stager installation records online data from the detector in the tape archive. The same installation serves data to the reconstruction farm and records its output data.

The DIM/Stager strategy has worked well, but providing robust service has taken longer than expected. The method automatically keeps the most popular data on disk and provides sharing of that data between different users. The ability in the DIM to fix some important data runs statically on disk for short periods has also been used. In the period of time since collisions started, about 40 TB of data have been written from the detector and from the reconstruction farm. During this time, resources of disk space and of tape drives have not been as adequate as planned for that volume of data, so that the appropriateness of the staging algorithms has been well tested and the algorithms tuned.

Future work on the DIM/Stager will introduce the ability to ration disk by user and physics group, the ability for user jobs writing datasets (defined later) to use a private DIM managed disk space, and monitoring tools for managers.

Data is organized hierarchically and this organization is described in a Data File Catalog (DFC). The DFC is a relational database, Oracle at Fermilab (and at non-Fermilab sites by network access), mSQL at some non-Fermilab sites. From the bottom up, the hierarchy is: events, runsections, files, filesets, and datasets. Runsections are the collection of events taken during about 30 seconds of detector time; runsections are used to record integrated luminosities. Datasets are collections of events of similar data content; primary datasets are defined by trigger bits set online (e.g. electron candidate with ET over 15 GeV) and contain reconstructed data. There is some overlap of events between primary datasets.

The basic functionality of the DFC was available early well before collisions started and was thoroughly tested in the mock data challenges. The description of the data obtained from the DFC is widely used in user jobs accessing the data and by the reconstruction farm in processing the raw data. A extensive web browser shows the DFC in many projections (as well as other databases online). The DFC shows the experimenter the progress of the data taking and of the reconstruction processing.

Work is underway to provide users the ability to extract secondary datasets from primary (or other secondary datasets). In the future, export and import of datasets and their DFC meta-data will be implemented.

Resource management (cpu time, tape drive use) has been accomplished, as planned, using the batch facility LSF. Only simple queue conditions have been implemented so far.

An extensive pair of IO modules[3] were written for the analysis framework, AC++[6]. These modules interface to the DFC and to the DIM/Stager. The DFC interface uses the package DataFileDB[7] which provides access either to the Oracle implementation of the DFC or to the mSQL implementation.

6 Analysis Systems

The central analysis system has to provide most of the analysis power required by the experiment and a modular design was chosen so that aging components can be replaced and new technologies can be integrated as they become available or mature. The minimum compute power required for the upcoming run of the CDF experiment was estimated, based on Run I experience, to be 3,000 SPECint95. The first large compute node was a 64 processor Origin 2000, commissioned at the end of 1999. It provides about one third of the estimated CPU requirement. The machine has a small amount of local SCSI disk space to hold a copy of the CDF reconstruction and analysis software. It has 100 Mbit/sec Ethernet point-to-point connections to the fileservers and a 1000 Mbit/sec Ethernet connection to the main network switch of CDF, a Cisco 6509. The second large node is due to be commissioned in August 2001. This is a 24 processor Sun F6800 which provides compute power similar to the Origin 2000.

To provide uniformity of login environment to users throughout the loosely coupled cluster of large compute nodes, CDF has two Network Appliances fileservers, an F740 and F780, with 300 GBytes of disk space each. Apart from the model, the configuration of the two fileservers is identical; however, they are not clustered in failover mode but are setup standalone. The first one serves the clusterwide user login area, the second fileserver provides the spool area. The fileservers have dedicated point-to-point network connections to all compute nodes in the central analysis cluster and a connection to the general purpose network for the desktop systems. The two fileservers have been running problem and maintenance free since their commissioning.

Fermilab selected Kerberos version 5 from MIT [8] for user authentication to avoid passwords being sent unencrypted over the network. The local password file is used for all other user account configuration information. All machines in the CDF Run II central analysis system require Kerberos login. Kerberos is also deployed on the desktop systems.

The design of the central analysis system is based on the idea of a tight data storage and CPU connection. The core of the data storage was planned to be a pool of over 20 TBytes of disk space. This disk space is to be used to cache data from tape and to assemble new datasets. Some of the datasets will be accessed very frequently and are expected to be kept disk resident for several months. The disk available now is 7 TB of which 3.6 TB is managed by the DIM. An extra 30 TB will be commissioned in the next two months. The disks use SCSI protocol over Fibre Channel[9], which then provides an efficient channel style protocol and a physical layer with few limitations. The network character of Fibre Channel is presently not used in the central analysis cluster.

The total data volume of CDF Run II is expected to be over 1 PByte. The media cost for this data volume is significant and the commodity tape technology has been chosen to reduce costs. Currently the tape technology in use is Sony AIT-2; an upgrade to AIT-3 is expected late this year.

CDF planned the central analysis system for what is called lights out operation, so that all data must be accessible without operator assistance. In 1998 CDF acquired an automated tape library from EMASS, now ADIC. The storage elements inside this AML/2 library can accomodate tape media of practically any form factor. In 1999 the library was upgraded to 4 quadro-towers. In this configuration the library can hold 1 PByte of data, assuming 50 GBytes 8 mm style media or similar density. The library has space for over 200 tape drives; currently 16 are connected to various compute and data serving nodes via parallel SCSI.

So far, about 800 tapes (50 GB) each have been written in the tape library. About 1% of tapes have given problems due to bad regions on the tapes as was expected from the evaluations by the Fermilab Serial Media Group. The problems with drives have been minimal. Maintenance of the AML/2 has needed close attention but this close attention has resulted in little downtime.

So far, about 800 tapes (50 GB) each have been written in the tape library. About 1% of tapes showed media problems and were replaced during the initial write. This fraction agrees with the expectation from the evaluations by the Fermilab Serial Media Group. Two of the tape drives failed within the first few tape read/writes and were replaced. The AML/2 tape library itself is working very well with almost no downtime other than for scheduled maintenance. Close attention needs to be paid during maintenace work on the library, however.

The CDF reconstruction farm is discussed in paper 1-005[10].

7 Summary

Since collisions started in April 2001, computing and analyzing activity has been intense. Rates of taking and analyzing data have been pushed to their maximum to test the robustness of the data processing and analyzing systems. Users have started to accustom themselves to the new way of describing the data and to the new access methods. The experiment has made a good beginning in its computing and analysis operations.

References

[1] F. Abe *et al.*, Nucl. Instr. and Meth. **A271** 387 (1988), see also `http://www-cdf.fnal.gov/pubcdf.html`.

[2] L. Buckley-Geer *et al.*, "Overview of the CDF Run II Data Handling System", CHEP2000, Paper 366,
http://www.fnal.gov/fermitools/abstracts/ftt/abstract.html,
http://www.fnal.gov/fermitools/abstracts/ocs/abstract.html.

[3] F. Ratnikov *et al.*, "User's Friendly Interface to the CDF Data Handling System", CHEP2001, Paper 4-027.

[4] R.Kennedy et al., "The CDF Run II event Data Model", CHEP2000 proceedings.

[5] S.Lammel and P.Hubbard, "The CDF Run II Disk Inventory Manager", CHEP2001, Paper 4-035.

[6] L. Sexton-Kennedy et al., "The Physical Design of the CDF Simulation and Reconstruction Software", CHEP2000 proceedings.

[7] D. Litvintsev *et al.*, "CDF Run II Data File Catalog", CHEP2001, Paper 4-037.

[8] M. Kaletka *et al.*, "Fermilab Strong Authentication Project", CHEP2000, Paper 306.

[9] X3T9.3/Project 755D/Rev 4.2, "Fibre Channel Physical and Signalling Interface (FC-PH)", Working draft, ANSI, October 8, 1993, The Fibre Channel Association, "Fibre Channel: Connection to the Future", ISBN: 1-878707-19-1.

[10] S. Wolbers *et al.*, "The CDF Run 2 Offline Computer Farms", CHEP2001, Paper 1-005.